AI4PEOPLE'S INSTITUTE REPORT TOWARDS AN ETHICS BY DESIGN APPROACH FOR AI

AI4PEOPLE INSTITUTE'S REPORT TOWARDS AN ETHICS BY DESIGN APPROACH FOR AI

How to design, develop and maintain responsible and just Artificial Intelligence (AI) systems that respect the fundamental rights, ethical, moral principles, and values of the European Union?

The report outlines in actionable detail an "Ethics by Design[1]" approach, which guides organizations, public or private, small, or large enterprises, in proactively designing, developing and maintaining Artificial Intelligence (AI) systems[2] in accordance with the laws, ethical, moral principles and public values that underpin the European Union. We also believe that the report offers an actionable blueprint, so that its execution can pave the way for improved future iterations.

The report provides a step-by-step description of an Ethics by Design process and criteria to assess lawfulness and ethical principles[3] to embed them into all AI system development lifecycle phases, by adopting pragmatic and operational Trustworthy AI system requirements (see footnote n. 1), ensuring these are developed in an accountable manner. The approach[4] is streamlined to integrate with existing operational processes and impact assessments[5], avoiding redundant work and additional expenses.

Thanks to the expertise of our team of leading experts from academia and business the report offers practical guidance[6] on how to implement each process phase, references to existing best practice methodologies and tools, along with recommendations for the European Union institutions on how best to foster and support the adoption of the Ethics by Design process.

Overall, the report argues that the Ethics by Design process is an effective method for both public and private organizations to achieve compliance when implementing AI systems. It is a valuable approach to ensure that AI systems do not only comply with the letter of the law, but are also responsible and just, with the spirit of ethics and trustworthiness. It is through this twofold compliance that AI systems can contribute to the promotion and protection of the fundamental rights of people and the common good of society. As argued in the "AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", the difference lies in playing well, not just by the rules: "Adopting an ethical approach to AI confers what we define here as a 'dual advantage'. On one side, ethics enables organisations to take advantage of the social value that AI enables. This is the advantage of being able to identify and leverage new opportunities that are socially acceptable or preferable. On the other side, ethics enables organisations to anticipate and avoid or at least minimise costly mistakes. With an analogy, it is the difference between playing according to the rules, and playing well, so that one may win the game"[7].

Michelangelo Baracchi Bonvicini
President, AI4People Institute

1 The term Ethics by Design (EbD) refers to an approach that aims to incorporate Trustworthy AI requirements that include regulatory requirements, ethical and moral principles into the design, development and monitoring process of an AI system. This definition is line with the definition and approach described in the EU Commission Paper title "Ethics By Design and Ethics of Use Approaches for Artificial Intelligence" Version 1.0 25 November 2021.

2 The term "AI System" is used in this document in accordance with the OECD definition provided in November 2023: "An AI system is a machine-based system that, for a given set of human-defined explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions influencing physical real or virtual environments." For more information visit OECD AI System definition.

3 See Guidance 7 "How to compile an Ethics Principles and Requirement Inventory?" and Guidance 8 "List of main Ethics Principles" of this document for more information.

4 The suggested approach is also in line with international standards such as ISO42001:2023.

5 Examples of impact assessments include but are not limited to Data Protection Impact Assessment ("DPIA"), Fundamental Right Impact Assessment ("FRIA").

6 The list of Appendix and Guidance can be found at the end of this document

7 Visit AI4People Institute website eismd.eu/ai4people/ to access and read the previous published report.

AI4PEOPLE INSTITUTE

IN BRIEF

AI4People Institute brings together academia, global businesses, civil society organizations and governments to consider the risks of AI and advocate for responsible and beneficial AI development and deployment.

AI4People Institute was launched in February 2018 as a pioneering research/policy project by Atomium-EISMD, Michelangelo Baracchi Bonvicini, Robert Madelin and Luciano Floridi to shape the debate on AI Ethics in the European Union and prompt European institutions to act quickly to stem future AI risks. Its action is at the origin of the regulatory process that led to the AI Act in Europe, the world's first AI regulation.

AI4People Institute focuses on various aspects related to AI, including ethics, policy, governance, privacy, inclusivity, and the responsible use of AI across different domains.

TABLE OF CONTENTS

1

INTRODUCTION

Artificial intelligence (AI) is a rapidly evolving and transformative technology that has the potential to bring significant benefits to various domains and sectors of society. However, AI also poses significant challenges and risks to the fundamental rights of people, the ethical and moral principles and public values that are at the basis of European laws. Therefore, it is essential to ensure that the development and deployment of AI systems are guided by a human-centric and value-based approach that respects the dignity, autonomy, privacy, equality, and many other values of the people affected by AI. At the community level, AI should not be used to undermine the integrity, independence and effectiveness of democratic institutions and processes, including the principle of separation of powers, respect for judicial independence, and access to justice.

One of the ways to achieve this goal is to apply an Ethics by Design (EbD) process.

The EbD process aims to incorporate Trustworthy AI requirements that include regulatory requirements, ethical and moral principles, considerations and requirements throughout the entire lifecycle of an AI system, from the design, development and monitoring process of an AI system.

The process assumes that ethics is not an afterthought or a constraint, but rather a proactive element that enhances the quality, compliance, reliability, acceptability, and trustworthiness of AI systems. It also aims to foster a culture of ethical awareness and responsibility among AI stakeholders, such as developers, providers, users, and regulators, and to promote a dialogue and collaboration among them on the ethical and value implications and impacts of AI[8].

[8] This refers to the notion of ‘cooperative responsibility’ as coined by Helberger et al. (2018) https://doi.org/10.1080/01972243.2017.1391913

Establishing a robust foundation for the adoption of an EbD process, mindset, and culture begins with a pivotal commitment from the CEO and top management.

It is from this top-down dedication that a cascading influence can permeate through the entire organizational structure, fostering a comprehensive approach to integrating ethical considerations into every aspect of operations. By prioritizing and championing ethical principles at the highest levels, leadership sets the tone for a corporate ethos that values integrity, accountability, and trustworthiness.

The EbD process is inspired by and aligned with the ethical and moral principles and public values that underpin the EU, as expressed in the EU Charter of Fundamental Rights, the EU Treaties, and EU legislation. It also draws from the ethical guidelines and recommendations issued by various European and international bodies and initiatives, such as the European Commission's High-Level Expert Group on Artificial Intelligence, the Council of Europe, the Organisation for Economic Co-operation and Development (OECD), the AI4People Institute Reports and UNESCO. Furthermore, the EbD process recognizes that ethical principles are contextual and system-specific and, therefore, requires a participatory value-elicitation approach which respects the moral expectations of people involved in the AI system design, deployment, use and maintenance.

The structure of this report is as follows. Chapter 1 describes the origins and the rationale of an Ethics by Design approach, and its relation to other similar concepts and approaches. Chapter 2 provides a step-by-step outline of the EbD process. Chapter 3 provides recommendations for the European Union institutions on how to foster and support the adoption of the EbD approach by the AI stakeholders in the EU. Chapter 4 concludes the report with final remarks. Additionally, a set of Appendices and Guidance are provided, to help the reader with some useful practical information for the application of the proposed process.

2

ORIGINS AND RATIONALE OF ETHICS BY DESIGN

Ethics by Design is not a novel or unique concept, but rather a synthesis and an adaptation of several existing concepts and approaches that have been developed and applied in different fields and contexts[9]. More specifically, EbD builds on and integrates the following approaches:

Value Based Engineering (VBE): integrates the Value Sensitive Design (VSD) tradition as well as the recent publication of the EbD standard ISO/IEEE 24748-7000. VBE and VSD are design approaches that aim to account for human values in a principled and comprehensive manner throughout the design process. It consists of several activities: conceptual, empirical, and technical investigations, which involve identifying, analysing, and addressing the values of the stakeholders and the system. VBE proposes to use moral frameworks, especially virtue ethics, to elicit peoples' value concerns as well as existing legal principles, both of which constitute "ethical value requirements" (EVRs) are then systematically translated into system requirements that constitute the EbD approach.

Data Protection by Design (DPbD) and **Privacy by Design (PbD[10])**: aim to embed privacy into the design and operation of information systems, networked infrastructure, and business practices. PbD was proposed by Ann Cavoukian in 2006[11] and it has been adopted and endorsed by various organisations and authorities, such as the European Commission, the OECD, and the International Conference of Data Protection and Privacy Commissioners. PbD consists of seven foundational principles: proactive not reactive, privacy as the default, privacy embedded into design, full functionality, end-to-end security, visibility and transparency, and respect for user privacy. It is one of the main sources of inspiration for the EbD process, as it provides a concrete and practical

[9] See also Brey, Philip, and Brandt Dainow. "Ethics by design for artificial intelligence." AI and Ethics (2023): 1-13.

[10] Rotenberg, Marc. "Artificial Intelligence and Democratic Values: The Role of Data Protection." Eur. Data Prot. L. Rev. 7 (2021): 496.

[11] Cavoukian, A. (2006). Creation of a Global Privacy Standard. Available at www.ipc.on.ca/images/Resources/gps.pd.

example of how to embed a specific value into the design and operation of systems.[12]

Responsible Research and Innovation (**RRI**): is a policy approach that aims to align research and innovation with the public values, needs, and expectations of society. RRI was proposed by the European Commission in 2018, and it has been implemented and supported by various programmes and projects, such as Horizon 2020, and the EU Framework Programme for Research and Innovation. RRI consists of six key dimensions: ethics, gender equality, governance, open access, public engagement, and science education. RRI also involves four key actors: researchers, policymakers, industry, and civil society. It can be considered as a valuable source of inspiration for the EbD Vprocess, as it provides a holistic and participatory approach to ensure the social and ethical acceptability of research and innovation.[13]

Ethical, Legal, and Social Implications (**ELSI**): is a research approach that aims to identify and address the ethical, legal, and social implications of emerging technologies, such as biotechnology, nanotechnology, and AI. ELSI was initiated by the US National Institutes of Health in the 1990s, and it has been adopted and expanded by various organisations and initiatives, such as the European Commission, the OECD, and the Human Genome Project. The approach involves conducting interdisciplinary and multidisciplinary research on the potential impacts and effects of emerging technologies on individuals, society, and the environment. It also involves engaging with various stakeholders and the public to inform and influence the development and regulation of emerging technologies. In Europe this perspective is discussed in terms of Ethical, Legal and Social Aspects (ELSA) (Fisher et al., 2006[14]), with more focus on establishing multi-stakeholder (ethical) boards, by involving citizen representatives, civil society organisations, policymakers, businesses, experts and other relevant actors. The EbD approach draws heavily on the foundation of ELSI/ELSA, as it provides a rigorous and comprehensive approach to assess and address the ethical, legal, and social implications of emerging technologies.[15]

[12] Cavoukian, 2011; Cavoukian and Jonas, 2012; Wright and De Hert, 2012.

[13] European Commission, 2013; Owen et al., 2012; Stilgoe et al., 2013

[14] Fisher, E., Mahajan, R.L. & Mitcham, C. (2006). Midstream modulation of technology: governance from within. In: Bulletin of Science, Technology & Society, 26(6), 485-496. https://doi.org/10.1177/0270467606295402

[15] Bostrom and Yudkowsky, 2014; Chadwick et al., 2014; Taddeo and Floridi, 2018

More in general the Ethics by Design approach fits in a longer tradition of multi-perspective formative exploration and review, used successfully in areas of industry and academia. This refers to collaborative techniques like Soft Systems Methodology (SSM), Formal System Model (FSM) and 'tool clinics' (Morton et al., 2013[16]). The conceptual basis for this type of socio-technological approach can be found in social constructivism and Science and Technology Studies (STS), as a reaction against technological determinism. It starts from the central concern that materiality, practice, and politics are necessarily entangled. Applying this perspective in a real-world context is typically indicated as Constructive Technology Assessment (CTA), with the objective to 'produce better technology in a better society'. This is done by intervening in the early stages of technology development based on the assessment of possible risks and opportunities that these technologies could have for society, and how to mitigate the former and improve the latter (Genus, 2006[17]).

EbD combines these approaches and is particularly useful when considering AI technologies. It provides an approach that is both principled and pragmatic, normative and empirical, proactive and reactive. Another key characteristic of the proposed process is to be flexible and adaptable, context-sensitive and stakeholder-oriented.

EbD is not a fixed or final approach, but rather a dynamic and evolving one, that can be modified and improved according to the feedback and the experience of the AI stakeholders, thus affording an approach born in Europe a more global footprint.

[16] Morton, A., Berendt, B., Gürses, S. & Pierson, J. (2013). 'Tool Clinics' – Embracing Multiple Perspectives in Privacy Research and Privacy-Sensitive Design (Chapter 4.3), Acquisti, Alessandro, Krontiris, Ioannis, Langheinrich, Marc & Sasse, Martina Angela (Eds.) 'My Life, Shared' - Trust and Privacy in the Age of Ubiquitous Experience Sharing (Dagstuhl Seminar 13312), Dagstuhl Reports, 3:7, Wadern: Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 96-104. https://doi.org/10.4230/DagRep.3.7.74

[17] Genus, A. (2006). Rethinking constructive technology assessment as democratic, reflective, discourse. In: Technological Forecasting and Social Change. 73 (1), 13–26. https://doi.org/10.1016/j.techfore.2005.06.009

3

ETHICS BY DESIGN PROCESS OVERVIEW

The Ethics by Design process guides the development of AI systems in accordance with the regulatory, ethical and moral principles and public values that underpin the European Union. The process is structured in such a way as to ensure that regulatory and ethical requirements are integrated into the design, development, and deployment of AI systems by small, medium, large, public and private organizations.

It is important to highlight that the process should be comprehensive and tailored to all types of AI systems, with content adjusted to the specific AI use case and possible internal or external changes. It should also be comparable to and integrated with existing processes, criteria and impact assessments (e.g., Data Protection Impact Assessment).

Phases

The process consists of the following six recommended phases:

Getting Started: Understanding organization context and foundation building

Phase 1: Understanding the AI System: Scoping and Specifications

Phase 2: Preliminary Impact Assessment

Phase 3: Trustworthy AI System Design

Phase 4: Implementation of Trustworthy AI System Design

Phase 5: Monitoring Trustworthy AI System Design

Each phase should be documented, traced, and formalized transparently and clearly, ensuring that the process is standardized and repeatable, with monitoring and control mechanisms in place to ensure quality, efficiency, and compliance with standards such as ISO / IEEE 24748-7000 and ISO 42001.[18]

The EbD follows a standard system development phases:

System Development Lifecycle Phases							
	Ideation	AI SYSTEM APPROVAL	Design	Build & Test	Deploy	AI SYSTEM GO-LIVE	Monitor & Update
EbD Phases[19]	Phase 1. Understanding the AI system: scoping and specification + Phase 2. Preliminary Impact Assessment		Phase 3. Trustworthy AI System Design	Phase 4. Implementation of Trustworthy AI System Design	Phase 4. Implementation of Trustworthy AI System Design		Phase 5. Monitoring of Trustworthy AI System Design

[18] See Appendix 1 "Ethics by Design Process Checklist" of this document to guide you through the process.

[19] Before the ideation of any new AI systems, it is recommended to perform a set of preliminary activities as described in the next section ("Getting Started – Understanding the organization context") as it sets the foundations and process to ensure responsible and trustworthy AI system design, development, monitoring and maintenance.

Getting Started: Understanding the organization context and foundation building

The cornerstone of ideating and deploying innovative technological solutions lies in understanding both the external and internal contexts within which an organization operates. By gaining insights into the broader organizational landscape, including its societal, cultural, and technological dimensions, organizations can lay down the essential foundations necessary for integrating ethical considerations seamlessly into the design and development of technological solutions.

Main Activities

· **Develop an AI Governance model** with comprehensive policies about the organizational processes, decision-making procedures and monitoring approaches for managing AI activities. To enable accountability, compliance and auditing, the AI Governance model should outline the roles and responsibilities of various stakeholders, including senior management and board members. A successful AI Governance model requires commitment from all levels in an organization: ensuring buy-in and commitment from senior management, therefore, is vital for AI Governance success.

· **Create a regulatory inventory** of the rules that apply to the organization[20], such as EU legislation, national legislation, international conventions, sectoral or domain-specific rules and standards, and internal organisations' rules and restrictions[21][22].

· **Identify the direct and indirect, internal and external, vulnerable and expert stakeholders** involved and impacted by the organization's AI projects, such as developers, providers, people interacting with it, beneficiaries, regulators and other affected parties to **understand their core values, concerns, expectations, interests, needs** and possible harms in relation to the AI realm[23].

[20] Visit IAPP website Global AI Legislation Tracker (iapp.org) for some information on existing and future legal frameworks.
[21] See Guidance 5 "Which are the main applicable AI Regulations?" of this document for more information.
[22] Visit AI Hub Standard website aistandardshub.org for a comprehensive list of international standards.
[23] See Guidance 2 "How to identify impacted stakeholders" of this document for additional information.

· **Create an ethical principles inventory**[24] that applies to the organization and its AI projects in general, including the ethical principles and values that underpin the European Union and internal organization codes and standards.

· **Translate** the ethical principles into operational **Trustworthy AI system requirements**[25] that can be technical, procedural and human oversight-oriented by leveraging and complementing the organization's existing system development requirements[26] and best practice standards and guidelines.

Expected Results

· A comprehensive AI Governance model including policies and guidelines for managing AI system development strategy, projects, and a clear outline of roles and responsibilities.

· The commitment of top executives to endorse and champion the AI Governance model.

· A comprehensive list of stakeholders as outlined in Guidance 2.

· A clear regulatory inventory that can be regularly reviewed, as outlined in Guidance 7.

· A list of ethical principles as outlined in Guidance 8.

· A list of clear Trustworthy AI System requirements as outlined in Guidance 10.

[24] See Guidance 7 "How to compile an Ethical Principles Inventory?" and Guidance 8 "List of main Ethics Principles" of this document for more information.
[25] See Guidance 10 "How to translate Ethics Principles into AI System Requirements" of this document for more information.
[26] Examples may include existing Security by Design requirements, Privacy by Design requirements, etc.

3.1 Phase 1 - Understanding the AI System: scoping and specifications

The objective of the first phase of the EbD process is to define the scope and gather the socio-technical specifications of the AI system context, a precondition to perform the preliminary and high-level analysis of the potential ethical impacts (see Phase 2).

Main Activities

1.1 Gather key contextual information on the AI system(s) use case from a business, social, technological and accountability perspective. This may include but is not limited to its ownership, purposes, values, scope, data, model, technology and expected outcomes[27].

1.2 Identify the applicable legal frameworks and regulations specific to the single AI system use case[28]. Use the regulatory inventory created in the previous phase to identify the main requirements by analysing the legal obligations of the AI system, such as compliance, accountability, liability, transparency, explainability, fairness, privacy, and data protection.

1.3 Identify the direct and indirect, internal and external stakeholders involved and impacted throughout the entire lifecycle of the specific AI system, by using the inventory developed in the previous phase to understand their core values in relation to the use case[29].

1.4 Document all the above in a clear, transparent and comprehensive manner, and communicate them to relevant stakeholders based on the need-to-know principle.

[27] See Guidance 1 "What information should I gather?" of this document for additional information.
[28] See Guidance 5 "Which are the main AI Regulations?" of this document for additional information.
[29] See Guidance 2 "How to identify impacted stakeholders" of this document for additional information.

To enable innovation while responsibly managing AI ethical concerns, it is recommended to develop an “AI Use Case Sandbox”[30] for experimentation and concept evaluation so that the viability and value to the organization can be tested responsibly on a minimal scope before the actual system design and development. This approach involves creating a temporary and controlled environment for the use case, where the AI system can be tested with appropriate legal and ethical safeguards and oversight to promptly manage risks.

Expected Results

- **AI System Use Case Scoping Form** using Guidance 1 as reference.

- **Stakeholder list** clarifying stakeholders’ roles, categories, short descriptions and requirements or expectations in relation to the AI System Use Case, using Guidance 2 as reference.

- **A Stakeholder Interaction Diagram** describing, for that specific use case, the interactions between stakeholders and the nature of such interactions.

[30] The recommended so called “AI Use Case Sandbox” approach differs from the standard “Regulatory Sandbox”. See Guidance 4 “AI Use Case Sandbox” of this document for additional information.

3.2 Phase 2 – Preliminary Impact Assessment

The objective of this phase is to perform a preliminary impact assessment of the specific AI system. The preliminary impact assessment involves several activities, such as conducting a preliminary and high-level analysis of the potential ethical issues, challenges, conflicts and impacts related to the AI system, defining the ethical principles applicable to the specific AI system, translating them into operational Trustworthy AI system requirements and verifying the lawfulness of the AI system. This phase ensures that the AI system is aligned with the legal, ethical and moral principles and public values that underpin the European Union. Moreover, it allows the identification of potential legal risks and liabilities that may arise from the design, development and deployment of the AI system, and to identify mitigating strategies to be adopted in Phase 5 – Implementation & Monitoring.

Main Activities

2.1. Assess the main ethical issues, challenges and conflicts associated with the specific AI system and define an inherent impact level. This activity is to be carried out while taking into consideration (a) the ethical principles inventory from the "Getting Started" phase, and (b) the key contextual information on the system scope, specifications and applicable regulatory requirements identified in Phase 1. It also includes (c) a stakeholder dialogue with the stakeholders identified in Phase 1 who interact with and are affected by the specific AI system. For the stakeholder dialogue, existing moral frameworks should be used to better understand the stakeholders' ethical concerns, values and virtues impacted by the system, as well as the public value implications of the AI system[31][32][33].

[31] See Guidance 3 "How to assess the main ethical issues, challenges and conflicts associated with the specific AI system?" of this document for additional information.

[32] At this stage it is recommended to perform a high-level reconnaissance of the potential impacts as opposed to deep dive impact assessments, which in turn is foreseen as a specific activity in Phase 3 Trustworthy AI System Design .

[33] The output of this activity may act as a trigger for other existing impact assessment processes (e.g., Data Protection Impact Assessment, Fundamental Right Impact Assessment, etc.).

2.2. Identify the ethical principles and values[34] applicable to the specific AI system and translate them into Trustworthy AI system requirements[35] that can be technical, procedural and human oversight-oriented by leveraging the organization's existing system development requirements[36] and best practice standards and guidelines. Following the identification of the main ethical issues, challenges, conflicts and impact level in the previous activity and taking into account the gathered insights from the stakeholder dialogue on the ethical concerns, values and virtues impacted by the system, in this activity ethical principles and values applicable to the specific AI system are identified from the ethical principles inventory ("Getting Started" phase) and translated into Trustworthy AI system requirements.

2.3. Determine the lawfulness of the AI system[37].

2.4. Document all the above in a clear, transparent and comprehensive manner, communicate them to relevant stakeholders based on the need-to-know principle.

Expected Results

· Analysis of potential ethical issues, challenges, conflicts and impacts associated with the specific AI system

· List of ethical principles and values and Trustworthy AI system requirements applicable to the specific AI system

· Determination of the lawfulness of the AI system, possible design changes, and mitigation measures to ensure continued compliance

[34] See Guidance 9 "How to assess applicability of Ethical Principles to AI systems" of this document for more information.
[35] See Guidance 10 "How to translate Ethics Principles into Trustworthy AI System Requirements" of this document for more information.
[36] Examples may include existing Security by Design requirements, Privacy by Design requirements, etc.
[37] See Guidance 6 "How to determine Lawfulness?" of this document for additional information.

3.3 Phase 3 – Trustworthy AI System Design

The third phase of the EbD process aims to design the AI system ethically through an extended ethical impact assessment, using as a starting point the preliminary system design, the preliminary assessment, and the applicable Trustworthy AI system requirements identified in Phase 2. The goal is 1) to clearly identify where the preliminary design does not meet the Trustworthy AI system requirements and 2) to define and prioritize the appropriate mitigating controls for the AI system working iteratively from the preliminary to the final design. This phase is crucial to ensure that the AI system is trustworthy.

Main Activities

3.1 Measure the impact levels of the AI system's features by assessing whether they meet each Trustworthy AI system requirement identified as applicable in Phase 2. As part of this activity, you can also estimate the severity of impact by examining the impact magnitude (how discernible would the impact on stakeholders be?) and the impact scale (how many stakeholders would be impacted?)[38].

3.2 Define the treatment strategy by specifying and prioritizing mitigating controls specific to the AI system, such as prevention, mitigation, enhancement, compensation, and monitoring of the ethical impacts and effects of the system[39].

3.3 Define the final Trustworthy AI system design and operations architecture[40] taking into account the selected technical, process and oversight control measures.

[38] See Guidance 11 "How to assess and estimate Ethical Impact ?" of this document for additional information.
[39] See Guidance 12 "How to select mitigating controls" of this document for additional information.
[40] See Figure 4 in Kreuzberger, Dominik, Niklas Kühl, and Sebastian Hirschl. "Machine learning operations (MLops): Overview, definition, and architecture." IEEE access (2023).

3.4 Document and trace the ethical impact assessment results, the treatment strategy, the resulting system design and operations architecture in a clear, transparent, and comprehensive manner, and communicate them to the relevant stakeholders.

Expected Results

· An assessment of the severity of ethical impacts as outlined in Guidance 11.

· A corresponding list of control measures, prioritized according to the severity level of each impact, with recorded implementation decisions as outlined in Guidance 12.

· A final Trustworthy AI system design and operations architecture incorporating the selected treatment strategies.

3.4 Phase 4 – Implementation & Monitoring of Trustworthy AI System design

The fourth and fifth final phases of the process have the objective of adopting, monitoring, and evaluating the implementation status of treatment measures identified in Phase 3. This represents a critical step in ensuring that AI systems operate ethically and responsibly. It is vital to ensure that the process and the AI system are effective, efficient, and adaptive, and that they respond to the changing needs and expectations of individuals and society. Moreover, this phase allows identification of the potential ethical innovations and learnings that may arise from the design, development, and deployment of the AI system, and to share and disseminate them in the AI community and beyond.

Phase 4 - Main Activities:

4.1 Implement the treatment measures identified in Phase 3.

4.2 Test the effectiveness of the implemented treatment measures.

4.3 Define and implement corrective actions needed to address any identified issues and also to prevent similar issues from occurring in the future.

4.4 Document the Trustworthy AI implementation and monitoring activities of the AI system in a clear, transparent, and comprehensive manner, and report them to the relevant stakeholders.

3.5 Phase 5 - Main Activities:

5.1 Establish mechanisms for monitoring the AI system's adherence to such treatment measures, and any emerging deviations from the Trustworthy AI System Design controls and evaluating the AI system's performance against these requirements.[41]

5.2 Design and activate feedback loop process to adjust the AI system based on the evaluation and monitoring activities.

5.3 Share the outcome of the entire process with interested and relevant stakeholders in the AI community.

Expected Results

- Trustworthy AI testing effectiveness results
- List of corrective actions
- List of monitoring activities and controls

[41] The monitoring activities may involve the use of automated tools and processes to track the system's behaviour, as well as regular reviews and audits to ensure that the system is operating in accordance with the established ethical principles.

4

CONCLUSION & RECOMMENDATIONS

In this report, we describe a high-level EbD process, which is designed to support AI system developers and providers who aim to adhere to the ethical principles and values central to the EU. We explain the origins, rationale, key stages, and components of the EbD process, and offer recommendations for how EU institutions can promote and support its adoption among AI actors within the EU.

The EbD process serves as a crucial framework to ensure that the design, development, and deployment of AI systems are driven by a human-centric and value-based approach that upholds the EU's fundamental rights and ethical principles. However, it is not a panacea or universally applicable solution; it demands ongoing, collaborative efforts from AI stakeholders to be implemented effectively and meaningfully. We provide specific strategies for EU institutions to encourage and facilitate the integration of the Ethics by Design framework among these stakeholders.

The recommendations are as follows:

- Promote the awareness and education of AI stakeholders on the EbD process, such as by providing information, guidance, training, and resources on the EbD process and its benefits and challenges.

- Encourage the participation and the engagement of AI stakeholders in the EbD process, such as by creating platforms, forums, networks, and events that facilitate dialogue, consultation, co-creation, and co-operation among the AI stakeholders on the EbD process and its outcomes and impacts.

- Support the implementation and the evaluation of the EbD process, such as by providing tools, methods, frameworks, and indicators that enable and facilitate application, assessment, improvement, and verification of the process and its results and effects.

- Ethics by Design emphasizes the importance of incorporating insights from a wide range of stakeholders during the development process. This includes the perspectives of groups that are frequently marginalized or excluded from public discourse due to economic, structural, or historical disadvantages. While their involvement is invaluable to developers, it can be particularly challenging for these groups. Consequently, it is crucial to ensure that their efforts are properly funded and compensated, and that their contributions are publicly recognized.

- Promote an evidence-based approach by collecting and sharing best practices of EbD processes and effective policies to support them. This could include the creation of a database with list of use cases and outputs of ethical assessments to help companies see previous examples of application in terms of considerations, stakeholders involved and mitigation actions applied.

- Ensure the coherence and consistency of the EbD process with the EU policies and initiatives on AI, such as by aligning and integrating the Ethics by Design process with the EU legal and ethical frameworks and regulations, the EU AI strategy and action plan, and the EU AI initiatives and projects.

- Provide access to legal consultants, researchers and experts to help organizations meet the applicable legal requirements.

APPENDIX 1 "ETHICS BY DESIGN PROCESS CHECKLIST"

Below is a high-level checklist to be support organizations in adopting the proposed EbD process:

Questions	Guidance	Useful Resources
1. Have you identified the purpose, scope, objectives, and expected outcomes of the AI system?	Guidance 1 - What information should I gather?	· IEEE 7000 Standard · Generalised methodology for ethical assessment of emerging technologies - zenodo.org
2. Have you identified the impacted stakeholders of the AI system?	Guidance 2 - How to identify impacted Stakeholders	· Framework for meaningful engagement of external stakeholders in AI development
3. Have you reflected on and traced the potential ethical concerns of the AI system deployment and possible use of AI Use Case Sandboxing?	Guidance 3 - What impacts should I consider? Guidance 4 - AI Use Case Sandbox	· Regulatory sandboxes in AI (OECD)
4. Have you identified the relevant legal frameworks and regulations that apply to the AI system?	Guidance 5 - Which are the main applicable AI Regulations?	· National AI policies & strategies - oecd.ai · Global AI Law and Policy Tracker - iapp.org
5. Have you applied these legal frameworks and regulations to determine its lawfulness?	Guidance 6 - How can I assess Lawfulness?	· How to ensure lawfulness in AI (UK ICO)

6. Have you identified the relevant ethical principles and values that apply to the organization and AI system?	Guidance 7 - How to compile an Ethical Principles and Requirements Inventory? Guidance 8 - List of main Ethical Principles	· EU Charter of Fundamental Rights - commission.europa.eu · Principles of the GDPR - commission.europa.eu · Ethics guidelines for trustworthy AI - digital-strategy.ec.europa.eu · ISO/IEC 42001:2023 - Artificial intelligence — Management system · IEEE 7000 Standard
7. Have you assessed the applicability of the ethical principles and public values to the AI system, based on its scope, specifications, and applicable regulatory requirements identified in Phases 1 and 2?	Guidance 9 - How to assess the applicability of Ethics Principles to AI Systems?	· Measuring adherence to AI ethics: a methodology for assessing adherence to ethical principles
8. Have you translated the ethical principles into operational Trustworthy AI system requirements?	Guidance 10 - How to translate Ethics Principles into AI Trustworthy AI system Requirements?	· Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment - europa.eu · IEEE 7000 Standard · Cybersecurity and privacy Criteria Catalogue for assurance and certification - truessec.eu

9. Have you evaluated the ethical impacts and effects of the AI system and assessed the fulfilment / deviation of its Trustworthy AI system design requirements?	Guidance 11 - How to assess and estimate the Ethical Impact?	· Algorithmic Impact Assessment tool · Impact Assessment Tool for ADM Systems in the Public Sector Toolkit · Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence | UNESCO · AI Ethics Impact Assessment Toolkit - Fujitsu Global · AI Ethics-SAQ - GSMA · Generalised methodology for ethical assessment of emerging technologies · Evolving to An Effective Algorithmic Impact · Good Work Algorithmic Impact Assessment
10. Have you defined and prioritized the appropriate treatment measures?	Guidance 12 - How to select mitigating controls?	
11. Have you implemented the Trustworthy AI system design requirements?		
12. Have you established mechanisms for monitoring the AI system's adherence to the requirements and evaluating its performance against them?		· NIST AI Risk Management Framework

13. Have you defined and implemented corrective actions needed to address any identified issues and prevent similar issues from occurring in the future?		· NIST AI Risk Management Framework
14. Have you shared the outcome of the entire process with interested and relevant stakeholders in the AI community?		· Agile Governance
15. Have you documented all the points above in a clear, transparent, and comprehensive manner, and reported them to the relevant stakeholders?		· AI Documentation (UK ICO)

APPENDIX 2 "ACRONYMS"

Acronym	Wording
AI	Artificial Intelligence
DPbD	Data Protection by Design
DPIA	Data Protection Impact Assessment
EbD	Ethics by Design
ELSA	Ethical, Legal and Social Aspects
ELSI	Ethical, Legal, and Social Implications
PbD	Privacy by Design
RRI	Responsible Research and Innovation
VBE	Value Based Engineering
VSD	Value Sensitive Design

GUIDANCE 1 "WHAT INFORMATION TO GATHER FOR THE AI SYSTEM USE CASE SYSTEM SCOPE AND SPECIFICATIONS?"

A comprehensive understanding of the AI Use Case including both the business, social and technical aspects is required for the exploration of its overall benefit and the potential ethical impacts.

The starting point of any AI use case development is the definition of appropriate accountability and responsibility. This should be done in accordance with the governance model prepared in the previous step. This ensures that it is always clear who is responsible for each aspect of system operations and machine decisions, especially in case of malfunction[42]. In practical terms, this means identifying and formalizing the project owner, manager, and other key roles.

A business case analysis helps to identify the benefits of the system and may point to insufficient problem-solution fit. System and data processing design, as well as interfaces and flows to external (sub-)systems, may have ethical consequences[43]. Therefore, it is recommended to carry out the following steps in the scope of activity 1.1 Gather key contextual information on the AI system(s) use case of Phase 1:

· **Create a structured AI Use Case Scoping Form** preferably based on a template to allow for inter-use case comparability (see example below)

· **Identify and fill in the minimum information required** to perform the preliminary impact analysis. Some of this information might become available only later in the design process and therefore the scoping and the impact analysis activity might need to be iteratively revisited. An update of the form is needed after the final system design.

[42] See for example, Active Responsibility gap in :Santoni de Sio, F., Mecacci, G. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philos. Technol. 34, 1057–1084 (2021). https://doi.org/10.1007/s13347-021-00450-x

[43] See for example: Suresh, Harini, and John Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle." Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. 2021. Models and data from external providers may have unknown, undesired properties.

The table below provides an example scoping form with some additional explanations. It also includes a practical example of a bank loaning system AI use case where appropriate (in italics):

Item	Contents
Use Case Ownership	
Project Owner / Sponsor	Insert name(s)
Project manager	Insert name
Development team	Insert name(s)
Other	
Use Case Overview	
Problem to solve / Business need	Objectively screen loans with speed.
Use Case Description	Screen loans with high accuracy based on past decisions as long as the interest rate is stable.
Category of Business Value	Increase efficiency / reduce cost / create new revenue / add value / meet customer expectations
Scope	
Users	Loan officers in Bank A
Geography	Democratic and capitalist nations or states
AI System Overview	
System name	Loan Review AI for Bank A

System scope	Describe the scope of the system · What (is given, input and output that the system subjects to?): Applicants' information, transaction history, and credit score to determine availability. · Who (is it intended for / uses it?): loan applicants (individuals) on Bank A's website · Where (in what type of environment or region is it used?): EU member states. · When (is it used?) 24 x 7. · How much: conventional default rate (0.05%), total latency within 1 min.
Expected functionality	Insert features expected during design and testing, for example: · to respond to the outcome within 5 min. in total to the applicant. · to recommend actions to redress for applicants when refuse. · to allow loan officer to intervene within 10 seconds to manage exceptional cases.
AI Model	· Insert high level information on the model and its origin – this may include: Internal or external Type: · Architecture:
Data	Insert high-level information relating to the data processed by the AI System – this may include: · Data required to train (if needed) and to operate the system · Data flows: (internal and external), data sources, data transformations, etc. · Data postprocessing and data augmentation strategies and responsible stakeholders · Interfaces with external parties (providers, partners, data annotators, etc.) · Format (structured / unstructured) vSize estimation

External APIs	Insert high-level information on the used APIs, for example: · APIs used · Input type content and size · Output type, content and size · API owner
User Interfaces	Insert high-level information on user interface specifics, for example with respect to accessibility · Developer interfaces · End-user interfaces
Operations	Insert high-level information on the operations architecture – this may include: · Intended MLOps (Machine Learning Operations) framework and architecture · Data drift observation and mitigation strategies
Cybersecurity & Data Protection	Insert high-level information on the cybersecurity and data protection aspects, if required, this may include: · Strategy · Special functional and technical measures
Other	Staffing availability and skillset (Ensure that appropriate resourcing and skillset been identified and allocated)
Other Stakeholders	**See Table Guidance 2**
Regulator Inventory	**See Table Guidance 5**

GUIDANCE 2 "HOW TO IDENTIFY IMPACTED STAKEHOLDERS?"

In line with the ISO 31000 and IEEE 7000 standards[44], a stakeholder is a person or organisation that can affect, be affected by, or perceive themselves to be affected by an AI system. Stakeholders can also be groups, communities, institutions, societies, future generations[45].

One method to comprehensively identify stakeholders is by applying categories. Several categories of stakeholders are discussed in the literature[46].

Categorization #1: Direct vs Indirect Stakeholders

- **Direct** stakeholders: interact directly with an AI system and are thus directly, positively or negatively, impacted by it.

- **Indirect** stakeholders: are indirectly, positively or negatively, impacted by the AI system due to their minimal interaction with it.

Categorization #2: Internal vs External Stakeholders

- **Internal** stakeholders: have an active role in an organization and can exercise more power in shaping an AI system, such as project team members.

- **External** stakeholders: are not directly involved in the organization such as end-users, suppliers, regulators, civil society or the public.

Categorization #3: Vulnerable and Expert Stakeholders

- **Vulnerable** stakeholders: these are usually highly impacted by an AI system but have little or no influence over it. Examples include children, the elderly, or people with protected characteristics.

- **Expert** stakeholders: comprise professionals with specialised knowledge in a certain field (technical, ethical, legal, sociological, etc.).

[45] Batya Friedman, and David G. Hendry. Value sensitive design: Shaping technology with moral imagination. MIT Press, 2019.

[46] See for instance Friedman and Hendry 2019, IEEE 7000, Philip Brey, Owen King, Philip Jansen, Brandt Dainow, Yasemin J. Erden, Rowena Rodrigues, Anais Resseguier, Marina Diez Rituerto, Tally Hatzakis, & Amal Matar. (2022). SIENNA D6.1: Generalised methodology for ethical assessment of emerging technologies (2.1). Zenodo. https://doi.org/10.5281/zenodo.7266895

It is recommended to maintain a broad perspective when exercising this activity and suggest the following activities:

· **Consult existing standards** and guidelines that provide standard roles. ISO/IEC 22989, for example, defined concepts and terminologies utilized in AI system development.

It defines the roles of stakeholders impacted by AI, such as the AI user or subject who is an organization or entity that uses AI products or services, or who is an organization or entity that is affected by AI systems.

· **Consult previous AI system scoping and specification** documentation to benchmark historical data (where feasible and available).

· **Consult with multiple and diverse members**, including those outside the AI system project management team, to gather additional input.

The process of stakeholder identification should produce two outputs:

1. Stakeholder list (recommended): a structured list and brief description of the people, groups and entities involved and impacted by the AI system, along with their main needs and expectations related to the AI system.

2. Stakeholder interaction diagram (nice to have): a diagram showing the interactions between the stakeholders listed in output 1.

Example of Output 1: Stakeholder List

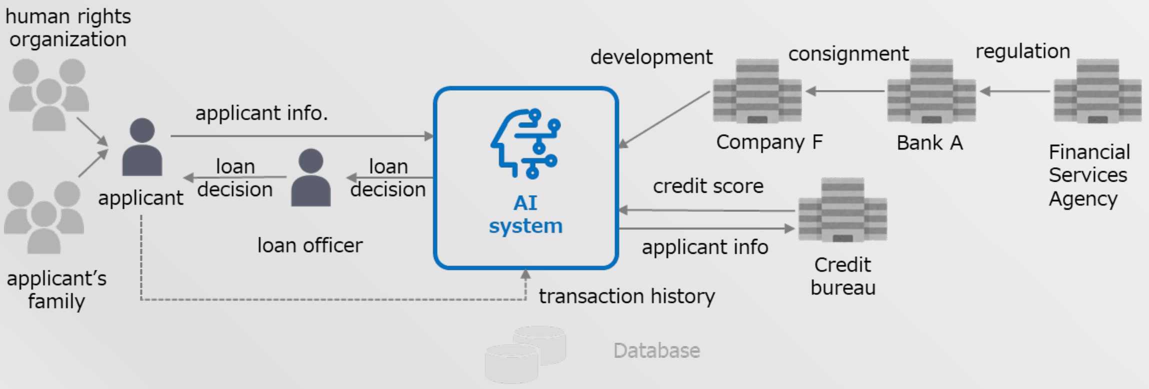
The table below provides a sample list of stakeholders, along with their requirements and expectations based on the example of a loan screening system:

<table>
<tr><th colspan="5">Use Case Scoping Form (Part of Guidance 2)</th></tr>
<tr><th colspan="5">Stakeholder</th></tr>
<tr><th>Role</th><th>Stakeholder</th><th>Description</th><th>Requirements and expectations</th><th>Category</th></tr>
<tr><td>AI Subject</td><td>Applicant</td><td>apply for the loan from a web form.</td><td>Prompt fair outcome or actions to redress when negative.</td><td>Direct, external</td></tr>
<tr><td>AI Subject</td><td>Applicant’s family</td><td>share a household with the applicant and be affected by the results.</td><td>Prompt fair outcome.</td><td>Indirect, external</td></tr>
<tr><td>AI Customer</td><td>Bank A</td><td>process applications using loan screening AI.</td><td>To increase profits or business by reducing workers or credit loss while increasing applicants.</td><td>Direct, internal</td></tr>
<tr><td>AI User</td><td>Loan officer</td><td>make decisions with final responsibility.</td><td>To maximize performance without losing the accuracy of decisions.</td><td>Direct, internal</td></tr>
<tr><td>AI Partner</td><td>Credit bureau</td><td>give the applicant’s credit score to Bank A.</td><td>Fair protection of personal and sensitive information like credit scores.</td><td>Indirect, external</td></tr>
<tr><td>AI Provider</td><td>Company F</td><td>develop loan screening AI for Bank A.</td><td>No particular expectations as long as the product satisfies Bank A’s requirements.</td><td>Direct, internal</td></tr>
</table>

Regulators	Financial service agency	create and enforce rules to which financial institutions must comply.	Automate decisions by AI to meet regulations.	Indirect, external
Relevant authorities	Human rights organization	protect human rights particularly of persons with disabilities, who are vulnerable, or from disadvantaged backgrounds.	Fair treatments and impacts of loan screening.	Indirect, external

Example of Output 2: Stakeholder Interaction Diagram

It is well known that visualization helps uncover concealed impacts. The image below provides a sample illustration of the interaction among the identified stakeholders based on the example of a loan screening system:



GUIDANCE 3 "HOW TO ASSESS THE MAIN ETHICAL ISSUES, CHALLENGES AND CONFLICTS ASSOCIATED WITH THE SPECIFIC AI SYSTEM?"

A key step of Phase 2 is the assessment of the main ethical issues, challenges and conflicts associated with the specific AI system (Phase 2.1). This task will help to define an inherent impact level. To complete this work, it is recommended to carry out the following activities:

1. Engage stakeholders, through a bottom-up approach, to examine people's core values and their concerns.

· This approach would benefit from a variety of interactions and engagement activities with stakeholders.

· To support this activity, ISO/IEEE 24748-7000 provides 3 questions to identify stakeholder values[47].

2. Create a list of ethical issues, challenges, and conflicts that should be considered when reflecting on the impacts of the AI system.

· The list should be based on the results of the stakeholder dialogue in the first activity, the key contextual information on the system scope, specifications and applicable regulatory requirements identified in Phase 1 and the ethical principles inventory developed in the "Getting Started" Phase.

[47] The 3 questions proposed by ISO/IEEE 24748-7000 (p. 39/49) stem from moral philosophy and are recommended in the following order:

Utilitarianism: What human, social, economic, or other values are affected positively or negatively, by the system?

Virtue Ethics: What are the negative implications of the system for the character and/or personality of direct and indirect stakeholders – that is, which virtue harms or vices could result if the system was implemented at scale?

Duty Ethics: What personal maxims or value priorities does the project team and organization see affected by the service that the project team members believe are so important that they want to preserve them in society?"

· If helpful, you could create a list of guiding questions to uncover ethical issues and challenges. A good starting point could be a quick review of existing frameworks such as the Ethics Guidelines for Trustworthy AI developed by the High-Level Expert Group on Artificial Intelligence, and the Assessment List for Trustworthy Artificial Intelligence (ALTAI).

3. Map the identified ethical issues, challenges and conflicts according to the relevant impact domain.

· Consider each ethical issue, challenge and conflict and identify the pertaining impact domain.

4. Review whether the AI system could negatively impact each identified domain, and the level of impact.

· The impact level could be based on broad categorisations (e.g., low-impact, medium-impact, high-impact), the traffic-light approach, or similar methodologies.

5. List potential treatment strategies for each impact identified.

This initial analysis aims to recognise early on whether some impacts are ethically impermissible, so that the AI system should be terminated, or its purpose modified. The results of the analysis can also be used to start identifying potential treatment strategies to address ethical issues and challenges.

The final output of this stage should be documented through the creation of a list or table. Below is an example that could be used as a reference. The result of this high-level analysis of potential impacts can inform some preliminary conclusions regarding the main ethical and legal concerns, and the technology and expertise required to develop the AI system and treat potential impact.

Below you can find a template table with some examples of questions to perform the preliminary assessment of ethical issues, challenges and conflicts. The table has been populated using as an example a loan screening system and does not aim to include a comprehensive set of guiding questions.

Preliminary assessment of ethical issues, challenges and conflicts				
Impact	**Guiding questions**	**Main ethical issues, challenges and conflicts**	**Impact level**	**Treatment**
Domain	**(examples)**			
Fundamental Rights	Does the AI system: · discriminate against groups of people? · pose a risk to children's rights? · pose a risk to personal data relating to individuals? · pose a risk to the freedom of expression?	It is unclear if the AI system discriminates against particular groups of people	Medium	Complete an assessment of bias in the training datasets and AI model, possible bias mitigation solutions
Human Agency and Oversight	Is the AI system designed to interact, guide or make decisions by human end-users that affect humans or society? Does the AI system risk creating human attachment, stimulating addictive behavior, or manipulating user behavior?	The intended use of the AI system requires some form of interaction and can guide human decisions	Medium	Establish clear processes to monitor the use of the AI system and effective human oversight

Societal and Environmental Well-being	Are there potential negative impacts of the AI system on the environment? Could the AI system have a negative impact on democracy?	The risks posed by the AI system to democracy and the environment are minimal	Low	No actions required
Assessment results				
Ethical concerns	The main concern is the negative impact that the AI system might have on Fundamental Rights and on Human Oversight.			
Legal concerns	No major legal concerns.			
Technology / expertise required	Technological solutions to assess and mitigate biases in the AI model, expertise in human oversight processes.			

GUIDANCE 4 "AI USE CASE SANDBOX"

As organizations seek to innovate utilizing AI they tend to be confronted with a high degree of uncertainty as a substantial portion of AI projects fail to reach production. Therefore, to instil confidence in the technical viability and potential value of a use case, we propose establishing an "AI Use Case Sandbox".

The proposed AI Use Case Sandbox provides a structured environment for exploration and assessment, facilitating informed decision-making and minimizing risks during the initial stages of experimentation and evaluation in the innovation process. This approach can be understood as a temporary, controlled, clearly bounded and secure virtual space where limited actors can access, collaborate and trial a use case.

It is further worth noting that the proposed "AI Use Case Sandbox" differs from AI regulatory sandboxes, which are set up under regulatory supervision and are utilized when a use case is expected to challenge or not be fully compliant with existing regulations. As such regulatory sandboxes commonly allow for waivers or exemptions of regulation to evolve the regulation itself. In contrast the proposed sandbox aims to abide by existing regulations and its focus is rather to enable rapid innovation towards building confidence in a use case's viability.

An "AI Use Case Sandbox" has the following characteristics:

1. It is **temporary** and should commonly only persist for several months

2. It is set up with **clear objectives** e.g., to prove the technical feasibility or value of the concept during the very early stages of the innovation process such as experimentation and concept evaluation.

3. It is bounded to a limited **scope** in terms for example of functionalities, type and quantity of impacted stakeholders, etc. required to evaluate its objective.

4. It is a safe and cyber-secure space with controlled access while respecting data privacy and internal company security and privacy by design guidelines and requirements.

5. During its existence it is **monitored** for ethical concerns.

When employing an “AI Use Case Sandbox” it is recommended to adopt the following guidance:

1. Identify the objective to be investigated within the sandbox. This could for example be a hypothesis to be tested or a value proposition to be confirmed.

2. Ensure that the use case under investigation shows no apparent risks of violating fundamental rights and does not fall under the high-risk or prohibited category of the EU AI Act or company policies.

3. Define the minimal scope and stakeholders required to evaluate the objective.

4. Let involved actors openly express any ethical concerns they may have, record and monitor the resulting perceived risk.

5. Ensure involved actors are aware of the “minimum recommended” ethical principles listed under Guidance 9.

6. Provision the sandbox as an insulated development environment including relevant means like storage, compute and memory capacities as well as tooling, libraries and frameworks for AI model development.

7. Provision the required data needed for the investigation while respecting data privacy and consent.

8. Ensure regular monitoring for ethical risks throughout the experimentation and duration of the Use Case Sandbox. Developing teams should further have access to subject matter experts in AI ethics in case of doubt or need of support.

9. In case obvious ethical risks with significant impact (as described in Guidance 12) or even violations are observed they should either be directly mitigated or the sandbox may be aborted and transitioned further to complete the EbD process.

10. Once the experiment or evaluation concludes the sandbox is closed and the EbD process is continued while respective findings on ethical risks are carried forward into design requirement.

As such the "AI Use Case Sandbox" provides a lightweight means to qualify use cases in the early stages of innovation, allowing to build confidence in the use case's chances of success while maintaining awareness of ethical concern through fostering awareness and dialogue of all actors.

GUIDANCE 5 "WHICH ARE THE MAIN AI REGULATIONS?"

Preparing an inventory for a particular AI system use case will require collaboration between the developers and expert advisors, so that the functionality and domain of operation of the project can be clearly and comprehensively discussed, and any relevant rules can be identified. As the context for each use case will be unique and the legal and regulatory environment for each industry is different, each organization must develop its own tailored regulatory inventory and keep it updated as rules change and international, national and sector-specific AI regulation expands and develops.

To conduct the activities 1.2 Regulatory Inventory and 2.2 Identify the applicable legal frameworks and regulations, it is recommended to carry out the following top-down activities:

1. Identify applicable external laws and regulations, such as:[48]

- International laws (based on AI system scope)[49]
- EU laws
- National laws
- Sector- or industry-specific regulations, standards, etc.

[48] Some useful starting points for this process include Information Commissioner's Office, How do we ensure lawfulness in AI? https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-lawfulness-in-ai/ ; Lifshitz, Lisa R; McMaster, Cameron. Legal and Ethics Checklist for AI Systems, Scitech Lawyer; Chicago Vol. 17, Iss. 1, (Fall 2020): 28-34.

[49] Some examples from outside the EU are noted in the table below. The Council of Europe has just adopted a Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, which may lead to further changes in the law governing AI in the future.

2. Identify applicable internal policies, codes and **standards** that are mandatory for the organization to adhere to, such as:

- Code of Conduct
- Ethics Policy
- Etc.

3. Identify applicable requirements for all identified external and internal laws and rules.[50] Questions that might be asked to surface possible legal or regulatory requirements include:

- **What does the system do?** For example, is it an online system? Is it consumer-facing? Does it allow the distribution of content? Does it deal with financial data?

- **How does it do it?** For example, does it process personal data? Does it use significant energy? Does it use AI to make decisions that have a legal or other significant effect on individuals?

- **Where will it do it?** The physical location where the system will be located or where its services will be accessed may bring new regulatory requirements into the frame. Local expertise may be required.

- **Is there sector-specific regulation?** Some sectors, such as the automotive industry, are already subject to specific regulations in some jurisdictions; many others are very likely to be introduced soon.

[50] In particular, consider whether the proposed system may be in the 'prohibited' or 'high-risk' categories set out in Articles 5 and 6 of the AI Act. Although this is an issue on which professional advice may be necessary, the summary infographic published by the International Association of Privacy Professionals may be useful as an initial reference point: https://iapp.org/resources/article/eu-ai-act-cheat-sheet/

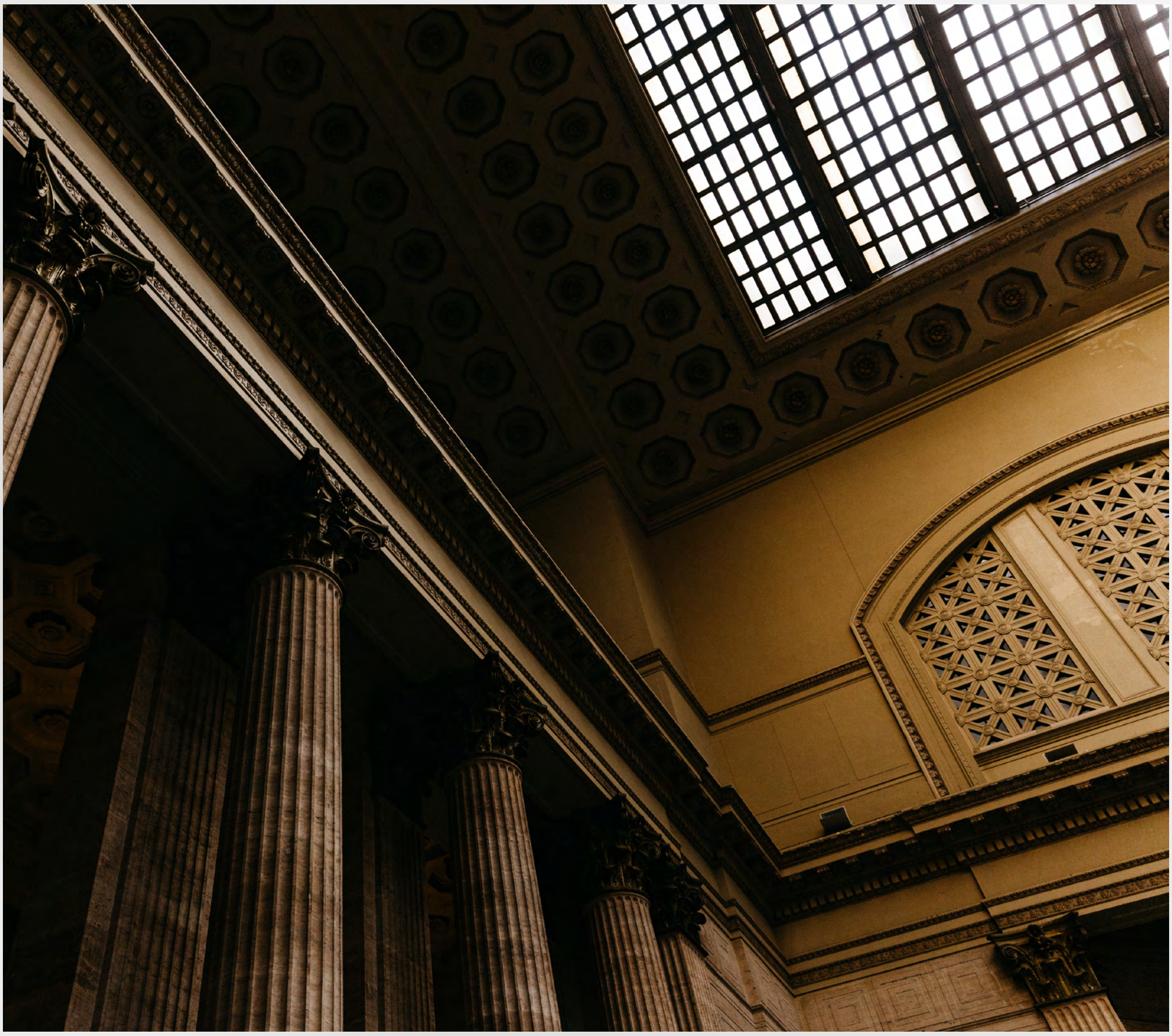
The table below illustrates an initial but not exhaustive list of external and internal rules applicable to the AI realm:

Laws	Applicability Guidance	Example of Key Requirements
European		
General Data Protection Regulation	Processing of 'personal data'	Article 35 (1) requires a Data Protection Impact Assessment when the processing of personal data is 'likely to result in a high risk to the rights and freedoms of natural persons'. Article 22 limits decisions 'based solely on automated processing, including profiling, which produces legal effects' or other significant effects for individuals.
AI Act	AI systems	Under Article 27, high-risk systems must undergo a Fundamental Rights Impact Assessment.
Digital Services Act	Online content and advertising	Article 34(1)(b) requires that very large online platforms (VLOPs) and very large online search engines (VLOSEs) carry out a risk assessment, which must include consideration of 'any actual or foreseeable negative effects for the exercise of fundamental rights'
Environmental Impact Assessment Directive	Public and private projects that are likely to have significant effects on the environment	Article 2 requires an Environmental Impact Assessment for such projects when they require a 'development consent'
Terrorist Content Online Regulation	Online content moderation	Article 3 requires that 'terrorist content' is taken down within one hour, which may require AI moderation
Cybersecurity Act	ICT products, services and processes	Voluntary cybersecurity certification

Digital Operational Resilience Act	Financial entities	Risk management, resilience testing, and oversight by European Supervisory Authorities
Clinical Trials Regulation	Clinical trials involving human subjects	Article 4 requires an ethical review and approval for such trials
Artificial Intelligence Liability Directive	AI products	Still being developed but is likely to create a rebuttable causal link between AI provider negligence and system output; note also the Utah Legal Personhood Amendments Act which denies legal personhood to AI
National / Local		
Artificial Intelligence	Colorado	SB205, the Colorado AI Act, will require developers and deployers to use reasonable care to protect consumers from known or reasonably foreseeable risks of algorithmic discrimination. Deployers will have to conduct impact assessments at least annually.
Algorithm transparency	France	The Law for a Digital Republic of 2016 requires access on demand to the main operational rules of a public administration algorithm
Algorithm transparency	Utah	The Artificial Intelligence Policy Act will require disclosure of the use of AI (for example, chatbots)
Intellectual property (IP) laws	All jurisdictions	Use of generative AI may require consideration of contested questions of IP rights; case law is developing and Ukraine has recently revised its copyright law to include a ‘sui generis right to non-original objects generated by a computer program’

Regulation of speech, such as hate speech or image-based abuse	All jurisdictions	Use of generative AI may require consideration of bias in underlying data and whether the AI can generate inappropriate or illegal images of real persons or children
Human resources	AI in hiring and promotion	New York City Local Law 144 requires independent bias audits of AI; Illinois Artificial Intelligence Video Interview Act 2022 requires notice and consent, and creates privacy and deletion rights; similar legislation is proposed in other US states
Industry / Sectors		
Healthcare	Varies by location	Future developments likely; note for example the US Food and Drug Administration Artificial Intelligence/ Machine Learning (AI-ML)-Based Software as a Medical Device (SaMD) Action Plan and California Bill AB 311, which may regulate 'automated decision tools' in healthcare
Financial services	Varies by location	Future developments likely; note for example the Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector; the EU Digital Operational Resilience Act may also be relevant
Insurance	Varies by location	Future developments likely; note for example Colorado Senate Bill 21-169, which restricts insurers' use of external data, and similar legislation is under consideration in other US states
Housing	Varies by location	Future developments likely

Automobile	Varies by location	Future developments likely; note EU Regulation (EU) 2022/1426 on uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles, combined with local laws such as German Autonomous Driving Act and Autonomous Vehicles Licensing and Operating Regulations
Aviation	Varies by location	Future developments likely



GUIDANCE 6 "HOW TO DETERMINE LAWFULNESS?"

Once the applicable laws and key requirements have been identified and assessed, the organisation needs to determine the lawfulness of the proposed AI system. This will again require the involvement of competent and expert legal stakeholders and advisors, such as general counsel.

It should be borne in mind that this is not a single-step process but ongoing and **iterative**: laws may change and new court decisions may provide new interpretations and understandings of how they should be applied. **Compliance is a journey** rather than a destination. It may also be that the legal analysis may not lead to a binary determination of lawful or not lawful but instead to an assessment of the relative risk of unlawfulness and heuristic predictions of the likely attitude of regulators. This open-ended conclusion will need to be considered in light of the ethical principles outlined above to arrive at a final decision on whether the system should be used or how it might be modified to achieve a more legally and ethically compliant outcome.

Nonetheless, an assessment of lawfulness will generally require the application of the logic of the laws and regulations to the system[51]. In principle, the use case(s) for the system should fall into one of three categories (although in practice, a lack of clarity may make these distinctions difficult to make and expert guidance may be required):

1. Does the law prohibit this use case?[52] These systems are simply unlawful and the analysis will be brief (in this case, do not proceed with the following steps). The underlying concept may need to be re-conceived from scratch so that it can developed lawfully.

[51] For a more detailed example of how this analysis should be conducted, see https://www.baden-wuerttemberg.datenschutz.de/legal-bases-in-data-protection-for-ai/

[52] Article 5 of the AI Act prohibits certain types of systems, such as predictive policing.

2. Does the law permit this use case, but subject to conditions?[53] If the system falls into this category, what are the specific conditions for this use case, industry and sector? Does the system comply with these? If not, can it be modified or re-designed to bring it into compliance without excessive cost?

3. Does the law not currently regulate this use case?[54] If the system falls into this category, it is lawful, although development should still be guided by EbD, particularly as this is likely to make the path to lawfulness smoother as requirements change in the future.

Updates to the law must be monitored on an ongoing basis to verify whether the legal status and categorisation of the system have changed.

[53] Article 6 and Annex III of the AI Act categorizes some types of systems (such as the evaluation of learning outcomes) as 'high risk'; these are permitted but are subject to certain requirements, particularly conformity assessment. Under Article 52, 'limited risk' systems are subject to transparency requirements. 'Minimal risk' systems have voluntary compliance obligations under Article 69.

[54] All AI systems have some regulatory obligations under the AI Act. As noted in Guidance 5, there may be national or sector-specific regulation that applies to the particular type of AI system that is being developed but whether or not it regulates the specific use case that is envisaged will be a context-specific question.

GUIDANCE 7 "HOW TO COMPILE AN ETHICAL PRINCIPLES INVENTORY?"

To create an ethical principles inventory as part of the "Getting Started" Phase, it is recommended to use the following four-level list of activities as guidance:

1. Identify the principles and values listed in the EU Charter of Fundamental Rights.

2. Identify EU regulations and ethics guidelines focusing on AI and the principles listed therein, such as:

- the AI Act
- the AI HLEG's Ethics Guidelines for Trustworthy AI
- the General Data Protection Regulation (GDPR).

3. Identify local regulations that are mandatory for the organisation to adhere to and the principles and values reflected in them.

4. Identify sectorial guidelines and the company's ethical principles and values, such as:

- the EIOPA's AI Governance Principles
- the Company's code of ethics or code of conduct.

The **first level** is the EU Charter of Fundamental Rights, which embodies the fundamental rights and freedoms that people in the EU enjoy in the midst of societal, technological and scientific changes. As a result, an ethical impact assessment for AI in Europe should explicitly include (but not be limited to) the principles and public values listed therein.

The **second level** consists of EU regulations and ethics guidelines that focus on AI, such as the AI Act, the AI HLEG's Ethics Guidelines for Trustworthy AI and the General Data Protection Regulation (GDPR). To ensure consistency with the AI Act, the ethical principles inventory should take into account the eleven organisational AI objectives specified in ISO 42001, Annex C.

As part of the **third level**, local regulations should also be considered to ensure alignment with the principles and values reflected in them (e.g., German National AI Strategy – Nationale Strategie für Künstliche Intelligenz, 2018).

The **fourth level** is comprised of sectorial guidelines (e.g., The European Insurance and Occupational Pensions Authority (EIOPA)'s AI Governance Principles) and the company's ethical principles and values contained in the company's code of ethics or code of conduct.

GUIDANCE 8 "LIST OF MAIN ETHICAL PRINCIPLES"

As each AI system has a specific context of use, scope and specifications, it is not possible to provide in advance a list of ethical principles applicable to all AI systems. Nevertheless, it is recommended to create an ethical principles inventory that applies to the organization and its AI projects in general (see Guidance 7) from which ethical principles applicable to each specific AI system can be further identified (Phase 2). There are various sources from which the ethical principles can be identified, such as:

- EU legal and ethical frameworks mentioned in Guidance 7
- Moral frameworks
- Stakeholder engagement
- Assessment of the specific use case
- Etc.

The following table provides a list of "minimum recommended"[55] ethical principles which can also be used as a basis for Guidance 9 "How to assess applicability of Ethical Principles to AI systems?"

Sources	Principles & Values
EU Charter	· Dignity · Justice · Freedoms · Equality · Solidarity · Citizen's rights
AI Act	· Proportionality in risk management · Human agency and oversight · Technical robustness and safety · Privacy and data governance · Transparency · Diversity, non-discrimination and fairness · Social and environmental well-being
GDPR	· Lawfulness, fairness and transparency · Purpose limitation · Data minimisation · Accuracy · Storage limitation · Integrity and confidentiality (security) · Accountability

[55] Other lists of ethical principles can be found in IEEE 7000, ISO/IEC 23894, ISO/IEC 38507, AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, OECD, etc.

AIHLEG	· Respect for human autonomy · Prevention of harm · Fairness · Explicability Key requirements · Human agency and oversight · Technical robustness and safety · Privacy and data governance · Transparency · Diversity, non-discrimination and fairness · Societal and environmental wellbeing · Accountability
ISO 42001	· Fairness · Security · Safety · Privacy · Robustness · Transparency and explainability · Accountability · Availability · Maintainability · Availability and quality of training and test data · AI expertise

GUIDANCE 9 "HOW TO ASSESS THE APPLICABILITY OF ETHICAL PRINCIPLES TO AI SYSTEMS?"

Any general list of ethical principles, such as those proposed in Guidance 9, should be considered as a starting point and should always be tailored to reflect the context of use of the specific AI system.

This is important for several reasons:

- First, different AI systems have different dimensions of impact.
- Second, an AI system may very often be used in more than one domain.
- Third, it is not uncommon for some principles to have a higher priority in certain domains than in others (e.g., finance vs transportation, health vs education).

Finally, the way principles are operationalised may differ between domains.

To assess the applicability of ethical principles to a concrete AI use case, it is recommended to conduct the following activities:

- Check the input of the analysis of potential ethical issues, challenges and conflicts identified in Phase 2.
- Consider the gathered insights from the stakeholder dialogue in Phase 2.
- Map the potential ethical issues and challenges arising from the deployment of the AI system identified in Phase 2 to corresponding ethical principles from the ethical principles inventory (Guidance 7 and 8).

The ISO/IEEE 24748-7000 standard provides practical guidance on how to comprehensively evaluate the applicability of ethical principles as well as to ensure effective stakeholder engagement. Additionally, the well-established methodology of value-sensitive design could also be helpful for the process of identification of ethical principles.[56]

[56] Refer to the following website for more information: https://standards.ieee.org/ieee/7000/6781/ Batya, and David G. Hendry. Value Sensitive Design, Shaping Technology with Moral Imagination. MIT Press, 2019; Friedman, Batya, Peter Kahn, and Alan Borning. "Value sensitive design: Theory and methods." University of Washington technical report 2, no. 8 (2002).

GUIDANCE 10 "HOW TO TRANSLATE ETHICS PRINCIPLES INTO TRUSTWORTHY AI SYSTEM DESIGN REQUIREMENTS?"

To fully operationalize ethical principles, it is important to translate them into concrete Trustworthy AI system design requirements. These requirements can be created by applying the ethical principles deemed applicable and the values identified by stakeholders to the specific use case under examination.

Trustworthy AI system design requirements can be formulated about the AI system or the stakeholders interacting with the AI system through its lifecycle (e.g., developers, deployers and end-users, as well as the broader society). Examples of this step can be found in:

- The translation of fundamental rights principles into the 7 requirements proposed by **the AI HLEG's Ethics Guidelines for Trustworthy Artificial Intelligence**[57], and the development of the **Assessment List for Trustworthy Artificial Intelligence**[58], aimed at assessing whether the AI system that is being developed, deployed, procured or used, adheres to these requirements.

- **Clause 9 of ISO/IEEE 24748-7000**[59], which outlines the process of translating prioritized values into concrete ethical value requirements, understood as "organizational or technical requirements catering to values that stakeholders and conceptual value analysis identified as relevant for the system" (p. 18).

It is important to note that these requirements can be satisfied not only through physical and functional features of the system design, but also through procedural and organizational features.

[57] Refer to the following website for more information: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[58] Refer to the following website for more information: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[59] Refer to the following website for more information: https://www.iso.org/standard/84893.html

The figure below provides an example of the transition from an ethical principle to concrete Trustworthy AI system design requirements in the context of a price optimization AI system developed by an insurance company.

Please note that this example does not intend to be comprehensive, but it is rather used to illustrate how ethical principles can be operationalized in practice.

Ethical Principle	Examples of Contextual Analysis	Examples of Trustworthy AI System Design Requirements
Fairness	The AI system's outputs do not contain any form of bias.	Training datasets are tested by the AI system developers to ensure there is no bias.
		AI system developers verify that the training data is diverse and sufficiently representative of end-users.
	Price optimization practices do not unfairly harm consumers.	Price optimization does not unfairly target vulnerable consumers to maximize their "willingness to pay".
		Mechanisms are in place to allow rapid flagging of discriminatory issues.

It is recommended that this step is reviewed with expert stakeholders. This will ensure that the list is clear and comprehensive and all relevant stakeholders to which they apply are explicitly included.

GUIDANCE 11 "HOW TO ASSESS AND ESTIMATE ETHICAL IMPACT?"

Ethical impact can be defined as any consequence of fulfilling/deviating from Trustworthy AI system design requirements[60]. To assess and estimate ethical impact, it is recommended to perform the three following steps:

1. Review if the Trustworthy AI system design requirements (pertaining to the AI system or the stakeholders interacting with it) are fulfilled/not fulfilled. The fulfilment of these requirements means that the ethical principles are protected and preserved, and this in turn can generate a positive impact on stakeholders. Conversely, the deviation from requirements can be the cause of a negative impact on stakeholders. To enhance transparency and support reporting and monitoring tasks, it is recommended that evidence is collected to show how each requirement is fulfilled.

2. Assess, for each impact identified, whether the impact is **positive or negative, short-term or long-term**, and the direct and indirect stakeholders that might be impacted.

3. Estimate the impact level using a well-defined rating system, for instance using a 'traffic light' approach or on labels such as high, medium and low. Although some quantitative approaches might be proposed to complete the estimation task, using quantitative methods could be challenging given the nature of AI systems and how they can ethically impact stakeholders. Even when qualitative methods are used, it is recommended that the estimation carefully evaluates the impact severity[61]. It is recommended to answer the following two questions to assess severity:

[60] This definition is in accordance with the definition of causal effect established by ISO 31000 standard on Risk Assessment: "an effect is a deviation from the expected. It can be positive, negative or both, and can address, create or result in opportunities and threats".

[61] Please note that the purpose of this activity is to assess the ethical impact, not the risks posed by the AI system. If, however, this guide is used to inform a pure risk management approach, in addition to severity also the likelihood of the risk would need to be assessed,

- Magnitude of impact: How discernible would the impact on stakeholders be?

This can be determined by reflecting on the duration of the impact (e.g., short-term, medium-term, long-term), the frequency of impact, and its reversibility.

- Scale of impact: How many stakeholders would be impacted?

It is important to acknowledge that estimating the severity of impact requires careful consideration concerning what is ethically permissible. Tensions might arise, for instance, in the case of AI systems that might have a slightly discernible negative impact on a significant among of people (moderate magnitude x large scale); or in the case of an AI system that might have a very discernible positive impact on the most, but a moderate negative impact on some minorities.

Useful resources

To complete Steps 1 and 2, it is possible to review or use existing toolkits. Below there is a list of toolkits that have been identified and reviewed as part of this report.

Impact Assessments developed for generic use by AI developers and providers	Impact Assessments developed for the Public Sector	Impact Assessments developed for the Workplace sector
SIENNA Project's Generalised methodology for ethical assessment of emerging technologies	UNESCO's Ethical Impact Assessment	Institute for the Future of Work's Good Work Algorithmic Impact Assessment
Fujitsu's AI Ethics Impact Assessment	Algorithm Watch's Impact Assessment Tool for ADM Systems in the Public Sector	
GSMA's AI Ethics Assessment	GovEx, the City and County of San Francisco, Harvard DataSmart, and Data Community DC's Ethics & Algorithms Toolkit	
Information Accountability Foundation and PWC's Evolving to An Effective Algorithmic Impact Assessment	Government of Canada's Algorithmic Impact Assessment Tool	

GUIDANCE 12 "HOW TO SELECT MITIGATING CONTROLS?"

The Trustworthy AI system design should be developed iteratively, starting with a preliminary design and then incorporating specific measures to address identified non-conformities with respect to Trustworthy AI system requirements. These measures constitute a treatment strategy tailored to the specific AI system.

Treatment strategies can employ various types of control measures, including monitoring, preventing, mitigating, or compensating for the ethical impacts and effects of the system. The minimum control measure for any identified non-conformance is monitoring. Strategies relying heavily on monitoring may require additional measures later in the system's life cycle if undesirable trends emerge. More advanced measures can promote thorough fulfilment of Trustworthy AI system requirements. Many examples of specific control measures can be found in the resources listed under 'Useful resources' in Guidance 11.

To support the selection of appropriate measures, we have further grouped them into 3 categories, each including activities that might be implemented in the various AI life cycle stages:

- Process control measures focusing on internal aspects of the organization responsible for the development or deployment and operation of the AI system. Typically, these measures describe required processes or process details.

- Human oversight control measures dealing with relations between AI operators and outside stakeholders and at the core with literal oversight of AI operation by human actors.

- Technical control measures focusing on implementational details of AI development and operations addressing actively the possible impacts.

Please note that, to select appropriate treatment measures, a careful analysis is recommended to consider how to balance the minimization of impact with the minimization of the resulting process and organizational overhead. As part of the analysis, for instance, the proportionality of treatment measures could be reviewed according to the impact level linked to not fulfilling each Trustworthy AI system requirement. Below we offer a non-exhaustive sample classification of possible available treatment strategies for negative impact on Liberty/Freedom/Human agency values based on available literature[62][63][64][65].

Category	List of measures
Process	· Internal review protocols · Internal audits · Internal release of model card report · Promoting training and education of workforce · Usage process documentation · Clear allocation of responsibility for each algorithmic decision to a human · Internal and external stakeholder engagement plan · Dialogue with impacted communities, groups and individuals · Involvement of humans whose work is being taken over · Check for criteria for moratorium on algorithm usage · Creation of a collaborative process with impacted communities, groups and individuals

[62] Stahl, Bernd Carsten, et al. "A systematic review of artificial intelligence impact assessments." Artificial Intelligence Review 56.11 (2023): 12799-12831.

[63] Refer to the following website for more information: https://ethicstoolkit.ai/

[64] Refer to the following website for more information: https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms

[65] Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence UNESCO, 2023

Human Oversight	· Enabling human interventions (organizational and technical level) · Review of machine and human overrides · Analysis of human overrides on bias · Release of public model card · Periodic external audits · Public performance monitoring · Public advisory group with decision-making authority on system use and system decisions · Human review of each algorithmic decision which is deemed relevant
Technical	· Performance of tests for release of a model card[66] · Periodic evaluation of algorithmic performance · Capture of relevant inputs, machine states and decisions in perpetuity · Ensuring that human adjudication results are fed into algorithm re-training · Deep review of algorithm usage and technical analysis of benefits vs non-algorithmic solution as decision guidance on further use

Once proportionate treatment measures have been selected, it is recommended that:

· Each decision regarding what treatment measure is enough to fulfil a Trustworthy AI system requirement is made explicit in a document.

· Such treatment measures are compared with existing company processes, oversight strategies and best practices in the development process. This comparison will promote the identification of new practices and suggested changes to existing policies to be implemented.

[66] A Model Card is a treatment strategy applicable for many ethical value clusters. It might contain the following information: Algorithm architecture, decision criteria for an algorithm vs. non-algorithmic solution, intended and unintended scope of algorithm usage, desired and undesired outcomes, plans to sunset the system, criteria for stopping the system in case of undesired outcomes. In order to create the Model Card tests covering the following aspects might need to be performed: Algorithm performance, robustness, privacy, bias and discrimination, safety of responses and decisions.

- The treatment measures and corresponding new practices and policy changes are reviewed to understand how they impact not only the technical execution of a project, but also on processes (e.g., with the creation of workflows around human oversight of models), and the structure of the organization in accordance with the organizations' strategy (e.g., in the case of a new need for oversight boards or specific positions).

Technical treatment measures are reviewed and clarified to ensure transparency and traceability, following best practices for software engineering, for example following the guidance of ISO 29148[67].

[67] Refer to the following website for more information: https://www.iso.org/standard/72089.html.

SCIENTIFIC COMMITTEE

– **Burkhard Schafer** | Professor of Computational Legal Theory, University of Edinburgh

– **Sergei Bobrovskyi** | Data Scientist, AI Platforms team, Airbus

– **Patrice Chazerand** | Former Director of Public Affairs, Digital Europe

– **Donald Combs** | Vice President and Dean, School of Health Professions, Eastern Virginia Medical School

– **Fausto Pedro Garcia Marquez** | Full Professor at Castilla-La Mancha University, Spain (UCLM)

– **Virginia Ghiara** | Principal Researcher - AI Ethics at Fujitsu

– **Hiroki Habuka** | Research Professor, Graduate School of Law, Kyoto University

– **Hiroya Inakoshi** | Project Director of Human-AI collaborative society project, Fujitsu

– **Rónán Kennedy** | Senior Lecturer at University of Galway

– **Jacob Livingston Slosser** | Assistant Professor of Law and Cognition, University of Copenhagen

– **Robert Madelin** | Former Director General, DG Connect, European Commission

– **Claudio Novelli** | Postdoctoral Research Fellow, Yale University

– **Cory Robinson** | Associate Professor in Communication Design & Information Systems, Linköping University

– **Sarah Spiekermann** | Business Informatics Professor, chairing the Institute for Information Systems & Society at Vienna University of Economics and Business (WU Vienna)

– **Jay Stoltzenberg** | Head of AI Services, Airbus

WORKING GROUP

– **Burkhard Schafer** | Professor of Computational Legal Theory, University of Edinburgh

– **Sergei Bobrovskyi** | Data Scientist, AI Platforms team, Airbus

– **Bianca De Teffé Erb** | Director & Data Ethics Leader, Deloitte (Rapporteur)

– **Virginia Ghiara** | Principal Researcher AI Ethics, Fujitsu

– **Hiroya Inakoshi** | Project Director of Human-AI collaborative society project, Fujitsu

– **Rónán Kennedy** | Senior Lecturer at University of Galway

– **Jay Stoltzenberg** | Head of AI Services, Airbus

– **Hristina Veljanova** | Project Assistant at Department of Law and IT, University of Graz

ACKNOWLEDGEMENTS

This report would not have been possible without the support of Atomium - European Institute for Science, Media and Democracy, FUJITSU Research of Europe and AIRBUS.

We are particularly grateful to the Working Group Members, the Scientific Committee and the staff of AI4People Institute (https://ai4people.org/ai4people-institute/) for their help.

4PEOPLE
INSTITUTE